A weighted combined effect measure for the analysis of a composite time-to-first-event endpoint with components of different clinical relevance

Rauch, Geraldine; Kunzmann, Kevin; Kieser, Meinhard; Wegscheider, Karl; Koenig, Jochem; Eulenburg, Christine

Published in:
Statistics in Medicine

DOI:
10.1002/sim.7531

Publication date:
2018

Received: 21 June 2016 | Revised: 15 September 2017 | Accepted: 26 September 2017

DOI: 10.1002/sim.7531

RESEARCH ARTICLE

WILEY Statistics in Medicine

A weighted combined effect measure for the analysis of a composite time-to-first-event endpoint with components of different clinical relevance

Geraldine Rauch[1,2,3] | Kevin Kunzmann[1] | Meinhard Kieser[1] | Karl Wegscheider[2] | Jochem König[4] | Christine Eulenburg[2,5]

[1]Institute of Medical Biometry and Informatics, University of Heidelberg, Im Neuenheimer Feld 130.3, 69120 Heidelberg, Germany

[2]Institute of Medical Biometry and Epidemiology, University Medical Center Hamburg-Eppendorf, Martinistr. 52, 20246 Hamburg, Germany

[3]Charité—Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Institute of Biometry and Clinical Epidemiology, Charitéplatz 1, 10117 Berlin, Germany

[4]Division of Pediatric Epidemiology, Institute of Medical Biostatistics, Epidemiology, and Informatics; University Medical Center of the Johannes Gutenberg University Mainz, Obere Zahlbacher Str. 69, 55131 Mainz, Germany

[5]Department for Epidemiology, University Medical Center Groningen, Hanzeplein 1 Groningen, 9713 GZ, Netherlands

Correspondence
Geraldine Rauch,
Charité—Universitätsmedizin Berlin, Institute of Medical Bioemtry and Clinical Epidemiology, Berlin, Germany.
Email: geraldine.rauch@charite.de

Funding information
German Research Foundation, Grant/Award Number: RA 2347/1-2

Composite endpoints combine several events within a single variable, which increases the number of expected events and is thereby meant to increase the power. However, the interpretation of results can be difficult as the observed effect for the composite does not necessarily reflect the effects for the components, which may be of different magnitude or even point in adverse directions. Moreover, in clinical applications, the event types are often of different clinical relevance, which also complicates the interpretation of the composite effect. The common effect measure for composite endpoints is the all-cause hazard ratio, which gives equal weight to all events irrespective of their type and clinical relevance. Thereby, the all-cause hazard within each group is given by the sum of the cause-specific hazards corresponding to the individual components. A natural extension of the standard all-cause hazard ratio can be defined by a "weighted all-cause hazard ratio" where the individual hazards for each component are multiplied with predefined relevance weighting factors. For the special case of equal weights across the components, the weighted all-cause hazard ratio then corresponds to the standard all-cause hazard ratio. To identify the cause-specific hazard of the individual components, any parametric survival model might be applied. The new weighted effect measure can be tested for deviations from the null hypothesis by means of a permutation test. In this work, we systematically compare the new weighted approach to the standard all-cause hazard ratio by theoretical considerations, Monte-Carlo simulations, and by means of a real clinical trial example.

KEYWORDS

clinical trials, composite endpoint, relevance weighting, time-to-event

1 | INTRODUCTION

Composite endpoints combine several events of interest within a single time-to-first-event variable. By combining several event types, the number of expected events is augmented, which decreases the variance and thereby increases the power. Moreover, when the clinical effect of interest cannot directly be captured by a unique event outcome, several event types

can be combined into a composite instead of formulating a multiple test problem for several primary event endpoints.[1,2] A composite endpoint can thus be interpreted as a surrogate for a time-to-event endpoint assessing the clinically most relevant type of event, eg, death. Generally, a surrogate should be highly correlated with the endpoint of primary interest in order to be clinically meaningful.

The common effect measure for composite endpoints is the all-cause hazard ratio which is tested by means of the logrank test or the Cox-model to include covariates.[3] This approach is based on counts of the total number of observed events and neglects the corresponding type of event. As a consequence, a major difficulty in the interpretation of clinical trials with composite endpoints is that the effect for the composite endpoint does not necessarily reflect the effects for the individual components.[4-6] In clinical application, the component effects are often of different magnitude or can even point in adverse directions. This problem becomes even more prominent, if the single endpoints forming the composite are of different clinical relevance. In this case, the composite endpoint does not define a meaningful surrogate and is not necessarily highly correlated to the endpoint of primary interest. Related guidelines therefore recommended not to combine endpoints of different clinical severity.[2,7,8] This claim, however, is unrealistic in practice. Fatal events (eg, cardiac death and death from any cause) usually define the most relevant events from a patient point of view. Therefore, these event types must be considered when defining the primary endpoint. On the other hand, fatal events are often relatively rare, that is, the expected number of events is low. Therefore, a combination with other event types within a composite would be helpful. However, any nonfatal event is clearly less relevant than death. For this reason, the current guideline recommendations can hardly be implemented in practice. As an alternative to a composite endpoint approach, multistate or, more specific, competing risk modeling allows to analyze multiple endpoints simultaneously.[9,10] Through additional assumptions on proportional baseline hazards or identical covariate effects across transitions, the power of a multistate model is increased compared to distinct Cox-models.[11] However, for the individual components of a composite endpoint these assumptions are usually not fulfilled and, as a consequence, multistate or competing risk models most often do not define a satisfactory option for this specific application. Therefore, methods are required, which ease the interpretation of composite endpoints even if the components are of different clinical relevance.

The CAPRICORN Trial[12] is an illustrative example to illustrate the pros and cons of composite endpoints in general and the impact of components with different clinical relevance in particular. This trial investigated the long-term efficacy of carvedilol regarding morbidity and mortality in patients with left ventricular dysfunction after acute myocardial infarction. In this multicenter, randomized, double blind and placebo-controlled trial, patients were randomly assigned to carvedilol or placebo in a 1:1 allocation. The originally planned primary endpoint was time to death from any cause. The accrual time was given by 24 months, the minimal follow-up by 3 months. Thus, the observational period is given by $[0; 27)$. During a masked interim analysis, the Data Safety Monitoring Board noted that the overall mortality rate was lower than anticipated and consequently, the study was likely to be seriously underpowered. Therefore, it was decided to change the primary endpoint to a composite time-to-first-event endpoint given as time to death or cardiovascular hospital admission. Thereby, it was intended to increase the number of events and thus augmenting the power of the trial. Clearly, the component death has a higher relevance than the component cardiovascular hospital admission. Therefore, it might be questioned whether a standard unweighted composite effect measure is suitable to judge the treatment performance.

The new composite primary endpoint was tested at a two-sided significance level of 0.0225, whereas a level of 0.0025 was saved to test the original primary endpoint. With this extremely small fraction of the global significance level, it is of course very unlikely to obtain a significant result and therefore this "saving" might be questioned. The trial was planned to detect a hazard ratio for death of 0.77 with a power of 0.9 at the one-sided significance level of 0.0225. Recruitment was planned to continue until 633 target events were observed which was assumed to require 1850 patients in total. After a mean follow-up of 1.3 years and inclusion of 1959 patients, the target number of events was reached for the new composite endpoint. The observed all-cause hazard ratio was given by $\hat{\theta}_{CE} = 0.92$ and failed significance by far ($p = .148 > .0225$), compare.[13] In contrast, the observed hazard ratio for death was given by $\hat{\theta}_{Death} = 0.77$ with a p-value of .0155. This shows that the contribution of the less relevant component "cardiovascular hospital admission" indeed had shrunken the overall composite effect. Note that for the CAPRICORN Trial,[13] two-sided significance levels and p-values were reported. For the sake of consistency within this manuscript, we report the corresponding one-sided significance levels and p-values here. With knowledge of the observed results in mind, the most reasonable endpoint in this study would have been a single endpoint given by "death." However, this knowledge was not available in the planning stage and during the ongoing study. From a planning perspective, it seemed reasonable to add the additional component "cardiovascular hospital admission" to potentially increase the power. However, at the same time, the influence of this new component should be limited as

"death" clearly defines the clinically more relevant endpoint. For this reason, the unweighted all-cause hazard ratio might be difficult to interpret in this case.

A possible approach solving this problem is to directly define a weighted composite effect measure, where the weights reflect the clinical relevance of the different components. By this, a treatment effect in the composite can no longer be caused by a large effect in a component of low clinical relevance. When introducing relevance weights for the components of the composite endpoint, the resulting modified surrogate endpoint is potentially clinically more meaningful and higher correlated to the endpoint of primary clinical relevance. Of course, even with a weighting approach, the treatment effect of the composite allows no generalization on the treatment effects of the individual components. However, this issue is less problematical if the composite really defines a clinically relevant endpoint, which truly reflects the effect of the intervention under investigation. For the specific application of the CAPRICORN Trial, a composite effect measure putting a higher weight on the component "death" than on the less relevant component "cardiovascular hospital admission" would be meaningful. By this, the negative impact of the less relevant component could be shrunken, which would ease the interpretation of the net effect.

Several authors have proposed weighting strategies for composite endpoints. Duc and Wolbers[14] recently presented a weighting approach for testing absolute risk differences that defines an alternative effect measure for binary composite endpoints. With respect to composite time-to-first-event endpoints, Pocock et al[15] and Buyse[16] proposed two similar concepts referred as the "win ratio" and the "proportion in favor of treatment," respectively, based on counting the number of pairwise comparisons with a more favorable outcome in the intervention group. Thereby, the components of lower importance only affect the comparison if the determination of the pairwise "winner" is not possible based on the component of primary importance. Bebu and Lachin[17] and Rauch et al[18] showed that these effect measures are highly dependent on the censoring distribution and are in particular influenced by competing risks. In addition, the influence of each individual component on the combined effect depends on the follow-up duration. Péron et al[19] proposed a modified proportion in favor of treatment correcting for bias introduced by uninformative censoring. However, when investigating several priority-ranked time-to-event outcomes, a competing risk situation arises, which corresponds to informative censoring. Moreover, the influence of the follow-up duration on the component weights does remain an unsolved problem. A further major disadvantage is that the computational effort for these effect measures becomes large if the underlying sample sizes increase. Despite these unfavorable properties, the win ratio and the proportion in favor of treatment are very present in the medical and in the statistical literature. Therefore, we will provide an exemplary simulation scenario to illustrate the difference between these approaches and the new method presented within this work. In addition, Lachin and Bebu[20] proposed a test based on a weighted average of the log-transformed component hazard ratios. In their original approach, the component hazard ratios are not given by the cause-specific hazard ratios but by the individual hazard ratio, which ignores the competing risk scenario. However, their approach can easily be adapted to the cause-specific hazard ratios to account for competing risks. A potential drawback of the latter method is that the weighted average of the log-transformed cause-specific hazards is not directly related to the common all-cause hazard ratio. Moreover, by weighting the log-transformed cause-specific hazard ratios, the number of events for each component does not influence the magnitude of the weighted effect measure but only its variance. By this, the weighted effect measure can be driven by a large effect in an individual component even if the component effect is highly variable due to a small event number. Finally, as the variance of the weighted average of the log-transformed cause-specific hazards is basically a weighted sum of the individual component variances, the variance increases with increasing number of components. In contrast, a composite endpoint combines several components with the aim to increase the number of events and thereby *reducing* the variance.

To overcome the above problems, we propose a new weighted composite effect measure defined as the ratio between the weighted averages of the cause-specific hazards for the two groups. This "weighted all-cause hazard ratio" defines a natural extension of the standard all-cause hazard ratio as the weights are assigned to the individual cause-specific hazards and not to the (log-transformed) cause-specific hazard ratios. In particular, if the weights for the individual components are equal, the new effect measure corresponds to the common all-cause hazard ratio. To estimate the weighted all-cause hazard ratio, estimators for the underlying cause-specific hazard are required, which can be obtained from any parametric survival model. In this work, we focus on the Weibull-model as the Weibull-distribution allows to model various shapes of event time distributions. The need for a parametric survival model might be seen as a potential drawback of our method. However, note that the commonly applied Cox-model to assess a composite endpoint is based on the assumption of proportional all-cause hazards. This is a very restrictive assumption as the hazards of the individual components sum up to the hazard of the composite endpoint, which thus cannot both be proportional between groups simultaneously, except for the case of equal baseline-hazards including the special

case of constant hazards (exponentially distributed event times). This strong proportional hazard assumption for the all-cause hazards is no longer required with our approach. Moreover, as the choice of the underlying parametric model to identify the individual cause-specific hazards is arbitrary, we feel that this does not define a too strong restriction.

This paper is organized as follows. In Section 2, the standard approach for quantifying and analyzing composite endpoints in terms of the all-cause hazard ratio and the logrank test will be introduced. In Section 3, we will shortly introduce the concept of the win ratio and the proportion in favor of treatment as these weighting approaches are currently broadly discussed in medical applications. Subsequently, the new weighted effect measure and a related test will be introduced in Section 4. Sections 5 and 6 illustrate and discuss the pros and cons of all approaches by theoretical considerations and by a simulation study based on a real clinical trial examples and on illustrative settings. We conclude with a discussion in Section 7. The R code implementing the calculations required for the illustrating examples is provided as Supporting Information to ease application in practice.

2 | STANDARD APPROACH FOR A COMPOSITE ENDPOINT

Throughout this work, a two-arm clinical trial comparing a new intervention to a control is considered where the index I denotes the intervention group and the index C is assigned to the control. The sample sizes in the intervention and the control group are assumed to be equally given by n for the sake of simplicity. Note, however, that the generalization of all presented methods to unequal group sizes is straightforward so this does not define a restriction. Moreover, it will be assumed in the following that the occurrence of an event is harmful, that means a lower number of events corresponds to a more favorable result. Furthermore, we will focus on the case where the aim is to demonstrate superiority of the new treatment and a one-sided test problem is formulated. In this work, a composite endpoint consisting of k components $EP_j, i = 1, \ldots, k$ is considered.

2.1 | Effect measure

The individual components can be expressed in a competing risk model with k competing events.[3] The components are usually parametrized via the cause-specific hazards for the two groups $\lambda^I_{EP_j}(t), \lambda^C_{EP_j}(t), j = 1, \ldots, k$.[3,21] The composite endpoint is then parametrized by the corresponding all-cause hazards given as the sum of the cause-specific hazards

$$\lambda^I_{CE}(t) = \sum_{j=1}^{k} \lambda^I_{EP_j}(t),$$

$$\lambda^C_{CE}(t) = \sum_{j=1}^{k} \lambda^C_{EP_j}(t).$$

Assuming proportional hazards, the all-cause hazard ratio which is defined as

$$\theta_{CE} := \frac{\lambda^I_{CE}(t)}{\lambda^C_{CE}(t)}, \tag{1}$$

is constant in time. The proportional hazard assumption is fulfilled if the hazard can be written in the common form of the Cox-model given as

$$\lambda(t) = \lambda_0(t) \cdot \exp(\theta \cdot X),$$

where $\lambda_0(t)$ is the baseline hazard and X is the binary covariate expressing the group allocation. For the all-cause hazard referring to the composite this means

$$\lambda_{CE}(t) = \lambda_{CE,0}(t) \cdot \exp(\theta_{CE} \cdot X), \tag{2}$$

and for the cause-specific hazards referring to the individual components the proportional hazard assumption implies that

$$\lambda_{EP_j}(t) = \lambda_{EP_j,0}(t) \cdot \exp(\theta_{EP_j} \cdot X), \quad j = 1, \ldots, k. \tag{3}$$

It is easily seen that (2) and (3) can only hold true simultaneously if the baseline hazards are equal across the components, that is,

$$\lambda_0 := \lambda_{CE,0}(t) = \lambda_{EP_j,0}(t), \quad j = 1, .., k. \tag{4}$$

Note that (4) implies that the instantaneous risk to experience an event at time t in a given treatment group is the same across all event types. However, as indicated above, fatal events usually occur only rarely whereas nonfatal surrogate event types are commonly much more frequent. Therefore, the common situation is that the baseline hazards for different event types are not all equal and, as a consequence, the proportional hazard assumption is not fulfilled for all components simultaneously. Despite the fact that (4) is usually not fulfilled, it is common practice to report the all-cause hazard ratio for the composite *and* the cause-specific hazard ratios for the components as constant parameters.

2.2 | Test problem and test statistic

The test hypotheses for a one-sided test problem formulated in terms of the all-cause hazard ratio are given as

$$H_0 : \theta_{CE} \geq 1 \quad \text{versus} \quad H_1 : \theta_{CE} < 1. \tag{5}$$

The standard test to assess the above hypotheses under proportional hazards is the logrank test. The score test version of the logrank test statistic is given as[22]

$$LR = \frac{\sum_{l=1}^{d} \left(I_l^I - \frac{N_l^I}{N_l^I + N_l^C} \right)}{\sqrt{\sum_{l=1}^{d} \frac{N_l^I N_l^C}{(N_l^I + N_l^C)^2}}}, \tag{6}$$

where d denotes the total number of observed events, N_l^l, N_l^C are the numbers of patients at risk just before the ith observed event ($l = 1, \ldots, d$) in the intervention and the control group, respectively, and I_l^l is an indicator variable, which equals 1 whenever the event occurred in the intervention group. The denominator corresponds to the common variance estimator from the Cox-model. Under the null hypothesis given in (5), the test statistic (6) is approximately standard normally distributed, where negative values of the test statistic favor the intervention.[23] Thus, the null hypothesis is rejected whenever $LR \leq -z_{1-\alpha}$, where $z_{1-\alpha}$ is the corresponding $(1-\alpha)$-quantile of the standard normal distribution and α is the one-sided significance level.

3 | CURRENT WEIGHTED APPROACHES FOR COMPOSITE ENDPOINTS

The "proportion in favor of treatment" is a new effect measure for prioritized outcomes which was introduced by Buyse.[16] First, the individual components are ranked according to their clinical relevance starting with the most relevant endpoint. Then, each patient in the intervention group is compared to each patient in the control group which results in a total of n^2 pairwise comparisons. The favor function $f(p_s)$ indicates the result of a comparison for a given pair $p_s, s = 1, \ldots, n^2$ as follows: First, the patient with the more favorable outcome with respect to the component of primary priority is determined. If no decision can be drawn due to censored observations, the comparison is based on the component of secondary priority and so on. If the patient from the intervention group is superior to the patient from the control group the favor function is set to $f(p_s) = 1$, whereas if the patient from the control group shows a better outcome then the favor function is set to $f(p_s) = -1$. If no decision for any of the endpoints is possible due to censored observations, the comparison between two patients is said to be "uninformative" and the favor function is set to $f(p_s) = 0$. The proportion in favor of treatment is then defined as

$$\Delta := E(f(p_s)) \in [-1; 1],$$

where positive values favors the intervention. Δ can be estimated as

$$\hat{\Delta} := \frac{1}{n^2} \sum_{i=1}^{n^2} f(p_s).$$

The one-sided test problem to be assessed in confirmatory analysis is then given by

$$H_0^{CE} : \Delta \leq 0, \quad \text{versus} \quad H_1^{CE} : \Delta > 0.$$

Buyse[16] proposed to assess these hypotheses by a permutation test. Note that this test strategy requires a substantial computational effort as calculating the point estimator $\hat{\Delta}$ already implies the investigation of n^2 pairwise comparisons and the permutation test requires to repeat this step for a large number of times. A more detailed mathematical description of this effect measure and the related test strategy was provided by Rauch et al.[18]

The win ratio introduced by Pocock[15] is similar to the proportion in favor of treatment presented above. In contrast to the proportion in favor of treatment, however, the win ratio was originally introduced for a matched-pair design where matching is done according to independent risk variables and thus there are only a total of n pairs to be compared. Using the favor function $f(p_s)$ with $s = 1, \dots, n$, the win ratio Ψ can be defined as follows:

$$\Psi := \frac{E\left(\sum_{s=1}^{n} 1_{\{f(p_s)=1\}}\right)}{E\left(\sum_{s=1}^{n} 1_{\{f(p_s)=-1\}}\right)} \in [0; \infty), \tag{7}$$

where $1_{\{*\}}$ is an indicator function, which indicates whether the comparison for the ith matched pair is positive or negative, respectively. The win ratio is thus defined as the expected number of "winners" divided by the number of "losers." It can be estimated by replacing the expected numbers by the observed frequencies. Large values of Ψ favor the intervention. As the matched-pair design is rather uncommon in clinical applications, we focus on the proportion in favor of treatment as the reference procedure within this work.

Several authors have shown that the proportion in favor of treatment and the win ratio highly depend on the underlying censoring distribution.[17,18] Moreover, Rauch et al[18] provided extensive exemplary settings to illustrate the strong dependence of this effect measure on the follow-up duration. The magnitude of the observed value $\hat{\Delta}$ is therefore difficult to interpret. In addition, it can easily be shown that the favor function f does not define a transitive relation between two observations meaning that if patient A wins over patient B, and patient B wins over patient C, this does not imply that patient A also wins over C. Rauch et al[18] therefore came to the conclusion that the above effect measures should not be used in the presence of censoring, which is the standard situation in a time-to-event setting. However, as the proportion in favor of treatment and the win ratio are so prominently discussed within the medical and statistical literature, we will include a comparative simulation study using the proportion in favor of treatment as an additional reference procedure.

4 | NEW WEIGHTED APPROACH FOR A COMPOSITE ENDPOINT

The idea of the weighted approach is to define a new composite effect measure that weights the influence of the individual components by their clinical relevance. In general, there are various ways to combine weighted components in a composite measure, compare, eg, the related works[15,16,20] discussed in Section 1. None of these approaches is directly related to the common all-cause hazard ratio defined in (1).

4.1 | Effect measure

In here, we propose the "weighted all-cause hazard ratio" defined as

$$\theta_{CE}^{w}(t) = \frac{\sum_{j=1}^{k} w_j \cdot \lambda_{EP_j}^{I}(t)}{\sum_{j=1}^{k} w_j \cdot \lambda_{EP_j}^{C}(t)}, \tag{8}$$

where $w_j \geq 0, j = 1, \dots, k$ are nonnegative weighting factors reflecting the clinical relevance of EP_j. For equal weights $w_1 = w_2 = \ldots = w_k$ the weighted all-cause hazard ratio corresponds to the standard all-cause hazard ratio. The weighted all-cause hazard ratio can also be interpreted as the standard all-cause hazard ratio based on modified component distributions that are parametrized via the modified cause-specific hazards

$$\tilde{\lambda}_{EP_j}(t) := w_j \cdot \lambda_{EP_j}(t), \quad j = 1, \dots, k.$$

This property is appealing, as the weighted effect measure thus defines a natural extension of the commonly applied all-cause hazard ratio. To ease the interpretation, the weighted all-cause hazard ratio can also be written as the standard hazard ratio multiplied by a weight-dependent factor

$$\theta^w_{CE}(t) = \frac{\sum_{j=1}^k w_j \cdot \lambda^I_{EP_j}(t)}{\sum_{j=1}^k w_j \cdot \lambda^C_{EP_j}(t)} = \frac{\sum_{j=1}^k w_j \frac{\lambda^I_{EP_j}(t)}{\lambda^I_{CE}(t)}}{\sum_{j=1}^k w_j \frac{\lambda^C_j(t)}{\lambda^C_{CE}(t)}} \cdot \theta_{CE}(t). \tag{9}$$

The weighted all-cause hazard ratio is a time-dependent effect measure as long as the cause-specific hazards have no common baseline-hazards. It can thus either be reported for a predefined reasonable time point t or, as an alternative, might be averaged over the complete observational period

$$\Theta^w_{CE}(f) := \frac{1}{f}\int_0^f \theta^w_{CE}(t)\mathrm{d}t, \tag{10}$$

where $[0, f]$ defines the observational interval.

Simple plug-in estimators for $\theta^w_{CE}(t)$ and $\Theta^w_{CE}(f)$ can be obtained by means of the estimators for the underlying cause-specific hazards. The latter can be obtained from any parametric survival model. In this work, we focus on the Weibull-model. Limitations with respect to the choice of the model will be discussed in Section 7. Let $\hat{\lambda}^I_{EP_j}$ and $\hat{\lambda}^C_{EP_j}$ denote the corresponding estimates of the hazard functions resulting from the corresponding Weibull-models, then an estimator for $\theta^w_{CE}(t)$ is given as

$$\hat{\theta}^w_{CE}(t) = \frac{\sum_{j=1}^k w_j \cdot \hat{\lambda}^I_{EP_j}(t)}{\sum_{j=1}^k w_j \cdot \hat{\lambda}^C_{EP_j}(t)}. \tag{11}$$

If the cause-specific hazards can reasonably be assumed to have the same baseline-hazard $\lambda_0(t)$, compare (2) to (4), an alternative nonparametric estimator for $\theta^w_{CE}(t)$ can be derived using the representation of the weighted all-cause hazard ratio given in (9). In case of equal baseline-hazards across the components, it holds that

$$\frac{\lambda^I_{EP_j}(t)}{\lambda^I_{CE}(t)} = \frac{\lambda_0(t)\cdot\exp(\theta_{EP_j})}{\lambda_0(t)\cdot\exp(\theta_{CE})} = \frac{\exp(\theta_{EP_j})}{\exp(\theta_{EP_j})} = \frac{\int_0^f \exp(\theta_{CE})\mathrm{d}t}{\int_0^f \exp(\theta_{CE})\mathrm{d}t} = \frac{\Lambda^I_{EP_j}(f)}{\Lambda^I_{CE}(f)}, \quad j = 1, \dots, k. \tag{12}$$

The latter expression is a ratio of the cumulative hazards over time which can be nonparametrically estimated by means of the corresponding Nelson-Aalen estimators $\hat{\Lambda}_{CE}(f), \hat{\Lambda}_{EP_j}(f), j = 1, \dots, k$. Thus, a nonparametric estimator for $\theta^w_{CE}(t)$ is given by

$$\tilde{\theta}^w_{CE}(t) = \frac{\sum_{j=1}^k w_j \frac{\hat{\Lambda}^I_{EP_j}(f)}{\hat{\Lambda}^I_{CE}(f)}}{\sum_{j=1}^k w_j \frac{\hat{\Lambda}^C_{EP_j}(f)}{\hat{\Lambda}^C_{CE}(f)}} \cdot \hat{\theta}_{CE}, \tag{13}$$

where $\hat{\theta}_{CE}$ defines the estimator for the all-cause hazard ratio from the common Cox-model. Generally, the components of a composite hazard usually do not correspond to the same baseline-hazards as the instantaneous risk to experience an event at time t differ between events of different type. Therefore, we focus on the parametric estimator (11) in the remainder of this work.

Both estimators (11) and (13) are consistent as the nominator and the denominator of (11) and (13) are both continuous functions of consistently estimated parameters which are thus also consistent by the continuous mapping theorem. As the denominator converges to a constant, applying Slutky's theorem yields the consistence, compare Lehmann.[24] A consistent estimator $\hat{\Theta}^w_{CE}(f)$ can be defined equivalently.

4.2 | Test hypotheses and test statistic

The test problem to be assessed in the confirmatory analysis in terms of the weighted all-cause hazard ratio are given as

$$H_0^{CE} : \theta^w_{CE}(t) \geq 1 \quad \text{versus} \quad H_1^{CE} : \theta^w_{CE}(t) < 1. \tag{14}$$

Equivalently, the test hypotheses can be formulated in terms of the integrated weighted all-cause hazard ratio

$$H_0^{CE} : \Theta^w_{CE}(f) \geq 1 \quad \text{versus} \quad H_1^{CE} : \Theta^w_{CE}(f) < 1. \tag{15}$$

A potential drawback of our approach is that the variance of the estimators $\hat{\theta}^w_{CE}(t)$ and $\hat{\Theta}^w_{CE}(f)$ cannot directly be deduced and therefore the asymptotic distributions are unknown. In order to overcome this problem, the above hypotheses can be assessed by means of a permutation test. The R source code for the corresponding permutation test is provided as Supporting Information, so application in practice is straightforward. Note that the knowledge of the underlying distribution is an important advantage of the weighted approach proposed by Lachin and Bebu[20] as the variance of the weighted average of the log-transformed cause-specific hazard ratios is a weighted sum of the variances of the cause-specific hazard ratios. However, according to the arguments presented in the introduction, our approach provides important interpretation advantages and therefore the choice of the weighted effect measure should not only be guided by the existence of a parametric statistical test.

4.3 | Considerations on the choice of weights

An important aspect of an adequate and meaningful analysis using a weighted composite effect measure is that the weights must be prespecified in the planning stage. In principle, the relevance weighting factors $w_j > 0, j = 1, \dots, k$ can be chosen completely freely without any restriction. The magnitude of these factors should thereby reflect the relative importance of the components to each other. Note that the influence of an individual component on the weighted effect measure becomes higher if the corresponding weight is large, but also if the underlying cause-specific hazard is large. Therefore, components referring to higher hazards are naturally up-weighted. A very small hazard caused by a low number of events will thus only influence the overall weighted effect if the relevance weighting factor is extremely large. In general, the largest weights should be assigned to the most harmful events such (e.g. fatal events). However, the specific clinical trial situation has always to be taken into account. As illustrated above, the weighted all-cause hazard ratio can be interpreted as the standard all-cause hazard ratio based on alternative component distributions which are parametrized via the modified cause-specific hazards. Recalling the planning situation of the CAPRICORN Trial,[12] we consider two types of events, namely, "death" and "cardiovascular hospital admission". From a clinical perspective, it seems reasonable to assign a higher weight to the more relevant component "death" and a lower weight to "cardiovascular hospital admission". As "death" is much more relevant than "cardiovascular hospital admission", we choose the following weights

$$w_{Death} = 0.9, w_{Hospital} = 0.1.$$

The question is now how the components weights modify the underlying true event time distributions. The reported observed results of the CAPRICORN Trial published in[13] comprised the all-cause hazard ratio for the composite endpoint, the cause-specific hazard ratio for death as well as the underlying observed event frequencies. From these results, we deduced the following underlying hazard assumptions and hazard ratios relaying on exponentially distributed event times

$$\begin{aligned}
\theta_{CE} &= \frac{0.0276}{0.0299} = 0.92,\\
\theta_{Death} &= \frac{0.008}{0.0104} = 0.769,\\
\theta_{Hospital} &= \frac{0.0196}{0.0195} = 1.005.
\end{aligned}$$

The estimated event time distributions of the composite and the individual components are thus given as

$$\begin{aligned}
S^I_{CE} &= 1 - \exp(-0.0276 \cdot t),\\
S^C_{CE} &= 1 - \exp(-0.0299 \cdot t),\\
S^I_{Death} &= 1 - \exp(-0.008 \cdot t),\\
S^C_{Death} &= 1 - \exp(-0.0104 \cdot t),\\
S^I_{Hospital} &= 1 - \exp(-0.0196 \cdot t),\\
S^C_{Hospital} &= 1 - \exp(-0.0195 \cdot t).
\end{aligned}$$

The modified event time distributions of the composite and the individual components are then given as

$$\begin{aligned}
S^I_{CE} &= 1 - \exp(-(0.9 \cdot 0.008 + 0.1 \cdot 0.0196) \cdot t) = 1 - \exp(-0.00916 \cdot t),\\
S^C_{CE} &= 1 - \exp(-(0.9 \cdot 0.0104 + 0.1 \cdot 0.0195) \cdot t) = 1 - \exp(-0.01131 \cdot t),\\
S^I_{Death} &= 1 - \exp(-0.9 \cdot 0.008 \cdot t) = 1 - \exp(-0.0072 \cdot t),\\
S^C_{Death} &= 1 - \exp(-0.9 \cdot 0.0104 \cdot t) = 1 - \exp(-0.00936 \cdot t),\\
S^I_{Hospital} &= 1 - \exp(-0.1 \cdot 0.0196 \cdot t) = 1 - \exp(-0.0196 \cdot t),\\
S^C_{Hospital} &= 1 - \exp(-0.1 \cdot 0.0195 \cdot t) = 1 - \exp(-0.0195 \cdot t).
\end{aligned}$$

The original and the modified event time distributions of the composite endpoint are also displayed in Figure 1.

It can be seen that the event time distribution curves for the intervention and the control group are relatively close for the unweighted approach whereas the curves are more distinguished for the weighted approach. As the weights are chosen to be smaller than 1, the weighted curves are generally shifted upwards. Note, however, that the weighted all-cause hazard ratio is a relative effect measure and hence the distance between the curves and not the position of the curves influences the overall effect. For the specific application of the CAPRICORN Trial, the proposed weights will thus increase the power compared to the standard all-cause hazard ratio approach as the more relevant component also corresponds to the higher effect.

However, the weighted all-cause hazard ratio can also result in a loss of power compared to the standard all-cause hazard ratio whenever the most relevant components show the smallest effects, which often is the case in medical applications. The rationale to use the weighted approach is thus not to increase power but to improve the clinical interpretation of the overall effect. In the latter case, the unweighted all-cause hazard ratio would suggest a too optimistic effect, which is importantly driven by a component of low clinical relevance. The RENAAL Trial[25,26] presented in Section 6 is an illustrative example for such a situation. These potentially conflicting interests in increasing the power and improving interpretation might impose the question how large the weights for the most relevant components can be chosen without too much loss in power. These considerations can be interpreted as a kind of "risk-benefit assessment." On the one hand, the aim is to provide a meaningful combined effect measure, which assigns higher weights to the more relevant components, on the other hand the weight must be chosen such that the less relevant components still contribute to the overall power to reach feasibility of the trial. Therefore, we recommend to choose some starting values for the weights based on external relevance criteria, eg, from a discussion with clinicians. Subsequently, the power related to the weighted all-cause hazard ratio under the given planning assumptions should be investigated by simulations. If the achieved power is too small, the weights are adjusted such that (1) the magnitude of the weights still reflects the clinical relevance and (2) the power loss is acceptable. This approach is illustrated for the RENAAL Trial[25,26] in Section 6. If no such weight constellation can be found, the sample size must be increased to guarantee both a powerful and a meaningful endpoint.

Generally, weighted combined effect measures are often criticized as the choice of the weights remains to a certain extent arbitrary. Although this criticism is in principle correct, it should also be kept in mind that the use of a composite time-to-first-event endpoint always refers to an implicit weighting of the components as the components, which correspond to a higher number of observed events have a higher influence on the combined net effect. Therefore, using the standard all-cause hazard ratio can also be criticized for an arbitrary implicit weighting of the components. It may there-

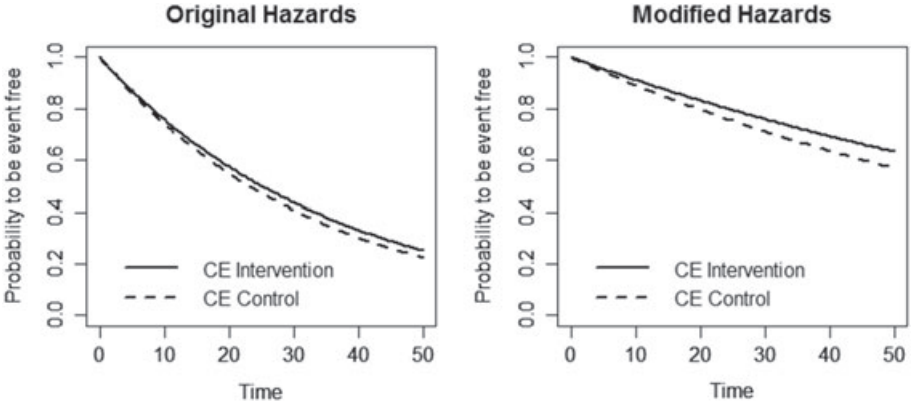

FIGURE 1 Event time distributions for the composite endpoint based on original hazards (left) and modified weighted hazards (right)

fore even be seen as a more objective strategy to assign predefined relevance weights instead of weighting exclusively by the amount of randomly observed event types. Independent of the chosen approach, whenever several time-to-event endpoints are combined within a single effect measure, there is no fully objective way to do so.

4.4 | Possible extensions

The weighted all-cause hazard ratio θ^w_{CE} can also be estimated based on covariate-adjusted estimates of the cause-specific hazards. It is also possible to include hazards estimates for the transition between a first and a second event or between any subsequent events. This has the advantage not to waive the information resulting from subsequent events occurring after the first. Unless for large cohort studies, the sample size of an interventional randomized, controlled trial is usually limited and thus the number of subsequent events is also limited, especially when differentiating between the different event types. As a consequence, the corresponding transition hazards will be small and therefore their impact on the magnitude of θ^w_{CE} is usually limited.

5 | COMPARISON OF THE DIFFERENT EFFECT MEASURES

5.1 | Comparison of the weighted and standard all-cause hazard ratio

To provide a systematic comparison between the standard and the new weighted approach, we begin by investigating the true treatment effects for given event time distributions and predefined component weights. For ease of representation, we consider a composite endpoint consisting of two components EP_1 and EP_2. We assume that the event times $T^I_{EP_1}, T^I_{EP_2}$ and $T^C_{EP_1}, T^C_{EP_2}$ are Weibull-distributed as

$$T^I_{EP_1} \sim \mathrm{W}\left(r^I_{EP_1}, s^I_{EP_1}\right), \quad T^C_{EP_1} \sim \mathrm{W}\left(r^C_{EP_1}, s^C_{EP_1}\right),$$
$$T^I_{EP_2} \sim \mathrm{W}\left(r^I_{EP_2}, s^I_{EP_2}\right), \quad T^C_{EP_2} \sim \mathrm{W}\left(r^C_{EP_2}, s^C_{EP_2}\right),$$

where $r^I_{EP_1}, r^C_{EP_1}, r^I_{EP_2}, r^C_{EP_2}$ denote the corresponding scale parameters and the shape parameters refer to $s^I_{EP_1}, s^C_{EP_1}, s^I_{EP_2}, s^C_{EP_2}$. The choice of Weibull-distributions is motivated by the fact, that they are able to flexibly model a variety of different distribution forms—in particular those distributions fulfilling the proportional hazards assumption and those who do not. To provide a wide range of distribution forms, the 5 parameter constellations provided in Table 1 will be considered. These 5 scenarios cover in particular the cases of constant hazards (scenarios 1 and 2), increasing hazards (scenario 3), and decreasing hazards (scenarios 4 and 5) over time. Moreover, the parameters of the Weibull-distributions are chosen such that the impact of the weights on the weighted cause-specific hazard ratio becomes obvious. Note that scenario 5 differs from scenario 4 in that the group-specific hazards of EP_2 are reversed. Thus, for scenario 5, the endpoints EP_1 and EP_2 show opposite effects.

The corresponding event time curves, the hazards, and the related (weighted) hazard ratios are displayed in Figures 2 and 3. The left part of Figures 2 and 3 shows the corresponding event time curves. The plot in the middle shows the underlying cause-specific hazards for EP_1 and EP_2 as functions in time. Finally, the plots on the right-hand side show the cause-specific hazard ratio for each endpoint, the standard all-cause hazard ratio that equals the weighted all-cause hazard ratio with common weights (here $w_1 = w_2 = 0.5$), and the weighted all-cause hazard ratios for weights $w_1 = 0.8, w_2 = 0.2$ and $w_1 = 0.2, w_2 = 0.8$, favoring either EP_1 or EP_2, respectively. Throughout this manuscript, we report

TABLE 1 Scale and shape parameters *r*and *s*of the Weibull-distributed event times

Scenario	$r^I_{EP_1}$	$s^I_{EP_1}$	$r^C_{EP_1}$	$s^C_{EP_1}$	$r^I_{EP_2}$	$s^I_{EP_2}$	$r^C_{EP_2}$	$s^C_{EP_2}$
1	0.9	1.0	1.0	1.0	0.4	1.0	0.8	1.0
2	0.1	1.0	0.6	1.0	0.15	1.0	0.2	1.0
3	0.3	1.4	0.8	1.0	0.15	1.3	0.2	1.0
4	0.3	0.8	0.8	1.0	0.1	0.9	0.15	1.0
5	0.3	0.8	0.8	1.0	0.15	1.0	0.1	0.9

the cause-specific hazard ratios as reference values for the component effects. Note, however, that the cause-specific hazard ratios are not directly related to the (weighted) all-cause hazard ratio and therefore a comparison of the composite effect to the component effects should be done with care.

Table 2 reports the different weighted all-cause hazard ratios $\theta^w_{CE}(t)$ at times $t = 1/3{\cdot}f,\ 2/3{\cdot}f,\ f$ and the averaged weighted all cause-hazard ratio $\Theta^w_{CE}(f)$ for an observational period of $f = 5$ years for different prespecified weights $w_1 = w_2 = 0.5$, $w_1 = 0.8, w_2 = 0.2$ and $w_1 = 0.2, w_2 = 0.8$. For the sake of comparison, the cause-specific hazard ratios $\theta_{EP_i}(5), i = 1, 2$, are also provided. Note that the weighted all-cause hazard ratio is equivalent to the standard all-cause hazard ratio for the case of equal weights across the components.

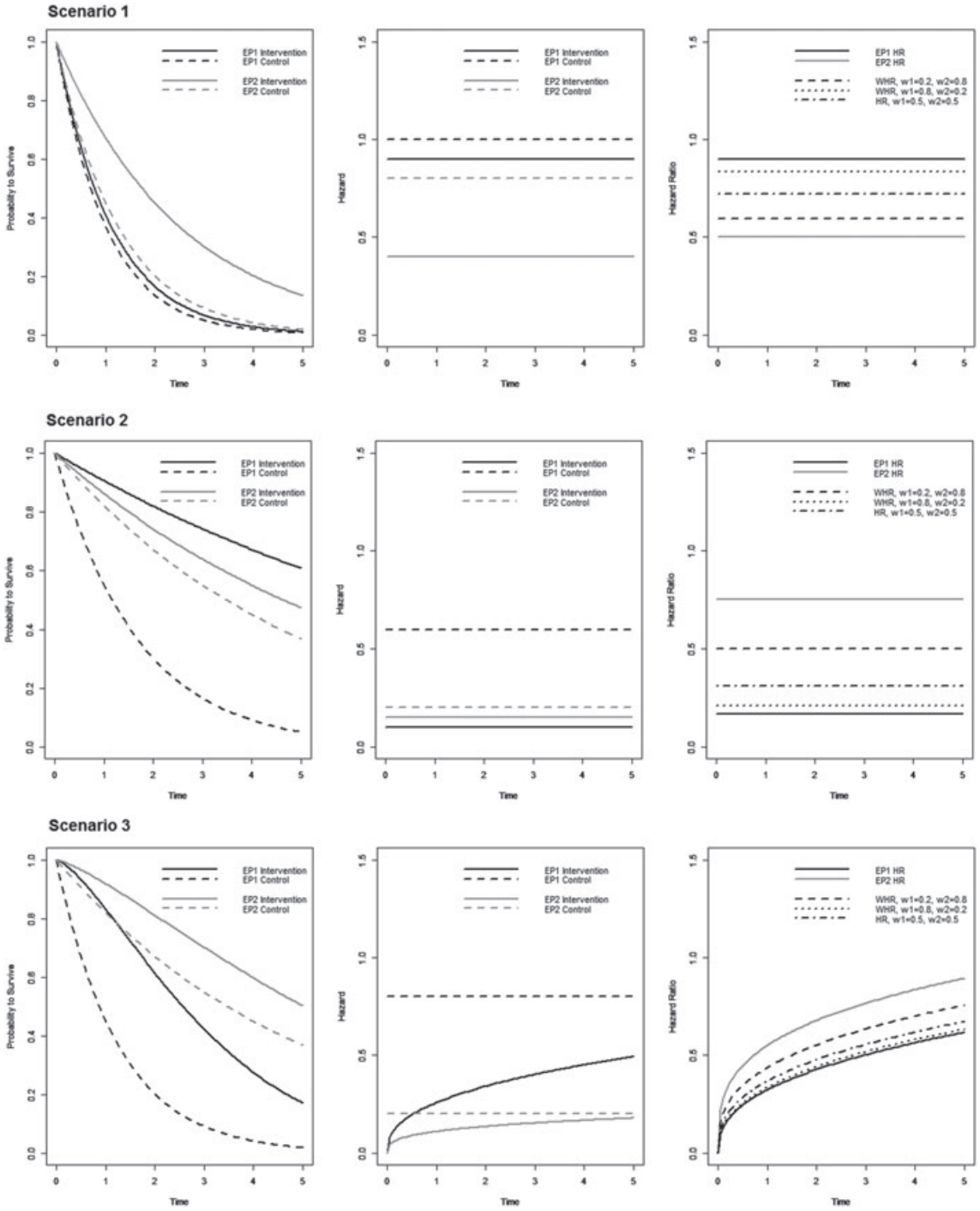

FIGURE 2 Event time distributions (left), corresponding cause-specific hazards (middle), and (weighted) hazard ratios (right) as functions in time for scenarios 1 to 3

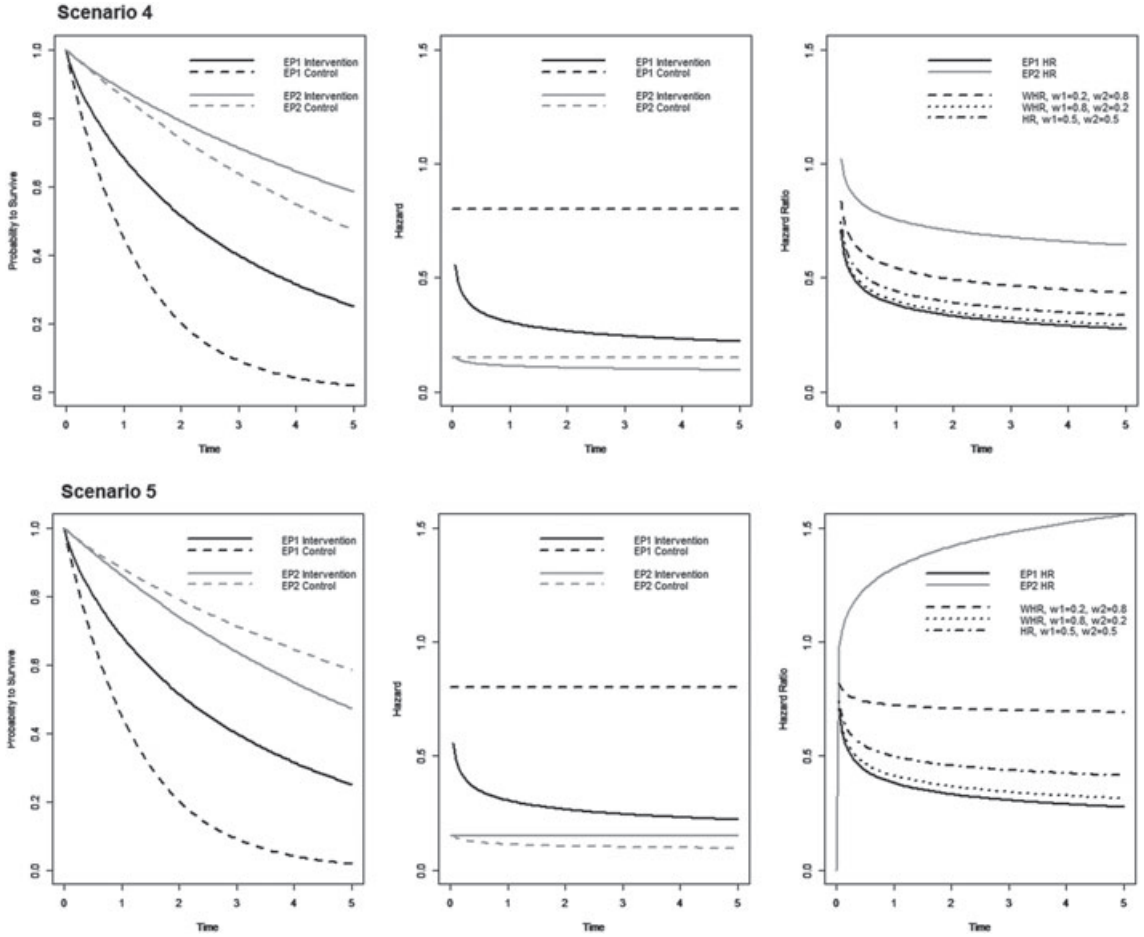

FIGURE 3 Event time distributions (left), corresponding cause-specific hazards (middle), and (weighted) hazard ratios (right) as functions in time for scenarios 4 and 5

TABLE 2 Comparison of different effect measures for scenarios 1 to 5

Scenario	w_1	w_2	$\theta_{EP_1}(5)$	$\theta_{EP_2}(5)$	$\theta^w_{CE}(1/3 \cdot 5)$	$\theta^w_{CE}(2/3 \cdot 5)$	$\theta^w_{CE}(5)$	$\Theta^w_{CE}(5)$
1	0.5	0.5	0.9	0.5	0.722	0.722	0.722	0.722
	0.8	0.2	0.9	0.5	0.833	0.833	0.833	0.833
	0.2	0.8	0.9	0.5	0.595	0.595	0.595	0.595
2	0.5	0.5	0.167	0.75	0.313	0.313	0.313	0.313
	0.8	0.2	0.167	0.75	0.212	0.212	0.212	0.212
	0.2	0.8	0.167	0.75	0.500	0.500	0.500	0.500
3	0.5	0.5	0.617	0.894	0.447	0.578	0.673	0.490
	0.8	0.2	0.617	0.894	0.412	0.541	0.634	0.456
	0.2	0.8	0.617	0.894	0.521	0.658	0.756	0.565
4	0.5	0.5	0.277	0.643	0.404	0.358	0.334	0.404
	0.8	0.2	0.277	0.643	0.361	0.317	0.293	0.362
	0.2	0.8	0.277	0.643	0.505	0.458	0.434	0.504
5	0.5	0.5	0.277	1.555	0.469	0.433	0.414	0.469
	0.8	0.2	0.277	1.555	0.379	0.336	0.314	0.380
	0.2	0.8	0.277	1.555	0.712	0.699	0.693	0.712

In scenario 1, the underlying hazards are all constant (exponentially distributed event times), therefore the dependence on time can be omitted. The cause-specific hazard ratios are given by $\theta_{EP_1} = 0.9$ and $\theta_{EP_2} = 0.5$. As expected, the all-cause hazard ratio $\theta_{CE} = 0.722$ lies in between these values. For weights $w_1 = 0.8, w_2 = 0.2$, the weighted all-cause hazard ratio

given as $\theta^w_{CE} = 0.833$ is closer to $\theta_{EP_1} = 0.9$, whereas for weights $w_1 = 0.2, w_2 = 0.8$ the weighted all-cause hazard ratio $\theta^w_{CE} = 0.595$ approaches $\theta_{EP_2} = 0.5$.

Similar considerations hold true for scenario 2 where the underlying hazards are also constant. However, for scenario 2 the underlying hazards for EP_1 are much larger than for EP_2. Therefore, assigning weights of $w_1 = 0.8, w_2 = 0.2$ brings $\theta^w_{CE} = 0.212$ even closer to $\theta_{EP_1} = 0.167$ whereas with inverse weighting $w_1 = 0.2, w_2 = 0.8$ the weighted effect $\theta^w_{CE} = 0.5$ is only slightly driven towards $\theta_{EP_2} = 0.75$.

The same holds true for scenario 3 where again the hazards for EP_1 are larger than for EP_2 and thus have a higher impact on the weighted effect. However, in scenario 3, the hazards in the intervention group are decreasing over time and therefore the (weighted) hazard ratios are no longer constant in time. The magnitude of the effect now depends on the time point where the averaged weighted all-cause hazard ratio shows the mean effect over time. However, the impact of the different weights remain similar, that is, EP_1 generally has a higher influence on the weighted all-cause hazard ratio due to the larger hazards.

In scenario 4, the hazards in the intervention groups are increasing in time. Again, the hazards for EP_1 are larger than for EP_2 and, as a consequence, EP_1 generally has a higher influence on the weighted all-cause hazard ratio.

Finally, scenario 5 is the same as scenario 4 except that the group-specific hazards for EP_2 are reversed. By this, EP_1 shows an effect in favor of the intervention (at $t = 5$ the effect is given by $\theta_{EP_1}(5) = 0.277$) whereas EP_2 shows an effect in favor of the control ($\theta_{EP_1}(5) = 1.555$). Interestingly, even with weights $w_1 = 0.2, w_2 = 0.8$, the averaged weighted all-cause hazard ratio given as $\Theta^w_{CE}(5) = 0.712$ is mainly driven by EP_1 and thus shows a clear effect in favor of the intervention. This again can be explained by the fact that the underlying hazards are larger.

The high impact of the magnitude of the underlying hazards on the weighted all-cause hazard ratio is an important difference to the weighted effect proposed by Lachin and Bebu[20] where the log-transformed cause-specific hazard ratios are weighted and thus the magnitude of the hazards has no influence. For real data sets, the magnitude of the hazards will influence the observed number of events. It seems intuitive to assign a lower impact to a component with only a few events, even if the corresponding cause-specific hazard ratio shows a large effect. If the number of events is low, the variability of the corresponding cause-specific effect is high and consequently the level of evidence is limited. Following this argumentation, it is a favorable property that components with lower hazards also have a lower impact on the weighted composite effect measure.

5.2 | Comparison of the weighted all-cause hazard ratio to the proportion in favor of treatment

The performance characteristics of the proportion in favor of treatment and the related win ratio approach for the specific application to time-to-event outcomes are already well understood, compare, eg, the works of Bebu and Lachin or of Rauch et al.[17,18] The main problems related to these approaches were highlighted in Section 1. Therefore, we generally do not recommend their use for time-to-event outcomes. However, the proportion in favor of treatment and the win ratio are widely discussed in medical applications. We therefore performed additional simulations to compare the proportion in favor of treatment to the weighted all-cause hazard ratio in order to better underline the differences.

Note, however, that the proportion in favor of treatment is calculated by comparing every patient from the intervention to every patient from the control group and that the significance is tested by means of a permutation test, compare Section 3. Assessing the power of this nonparametric test again requires multiple repetitions of these steps. Therefore, the computational effort to evaluate the power of the proportion in favor of treatment test is huge. As a consequence, we could only investigated an artificial scenario with a small sample size and a medium number of permutation and simulation runs. The obtained power values are therefore limited in precision. Note that the computational effort of the proportion in favor of treatment test is a one of the drawbacks of this approach.

We consider a composite endpoint consisting of two components with different clinical relevance, where EP_1 defines the more relevant endpoint and EP_2 the less relevant component. The predefined weights for the weighted all-cause hazard ratio are given as

$$w_1 = 0.8, \quad w_2 = 0.2.$$

We further assume exponentially distributed event times for each component with the following underlying hazards

$$\theta_{CE} = \frac{0.4}{0.2} = 2.0,$$
$$\theta_{EP_1} = \frac{0.03}{0.06} = 0.5,$$
$$\theta_{EP_2} = \frac{0.37}{0.14} = 2.64.$$

TABLE 3 Simulated all-cause hazard ratio, weighted all-cause hazard ratio and proportion in favor of treatment with related power values

Accrual	Minimal FU	$\hat{\theta}^{w}_{CE}$ (Power)	$\hat{\theta}_{CE}$ (Power)	$\hat{\theta}_{EP_1}$ (Power)	$\hat{\theta}_{EP_2}$ (Power)	$\hat{\Delta}$ (Power)
3	5	1.44 (0.01)	2.09 (0.00)	0.93 (0.03)	2.48 (0.00)	0.14 (0.19)
2	10	1.44 (0.00)	2.06 (0.00)	1.03 (0.02)	2.28 (0.00)	0.21 (0.37)

Thus, there exist a large negative effect for the composite endpoint due to an even larger negative effect in the less relevant component EP_2 whereas the more important component EP_1 shows a strong positive effect.

For these hazard assumptions, we randomly generated 1000 random samples with 40 patients per group. Thereby, uniform patient accrual was simulated with a fixed minimal follow-up. Two scenarios were considered to assess the impact of the individual follow-up distribution on the different effect estimators: An accrual duration of 3 months with an additional minimal follow-up of 5 months and an accrual duration of 2 months with a follow-up of 10 months. For both scenarios, the same starting seed was used for simulations to make the results of both scenarios more directly comparable. For each sample, the proportion in favor of treatment was estimated as described in Section 3. The permutation test was performed by randomly reassigning the group affiliations to the original complete data set and subsequently calculating the point estimator $\hat{\Delta}$ as described above. We performed 500 repetitions of permutations to obtain the null distribution of the estimator $\hat{\Delta}$. The corresponding hazard estimates within the two groups were obtained by fitting Weibull-models with scale parameters fixed to 1 using the function `survreg` of the `R` package `survival`. The plug-in estimator $\hat{\theta}^{w}_{CE}$ was obtained by substituting the resulting hazard estimates. The corresponding permutation test was performed by randomly reassigning the group affiliations to the original complete data set and subsequently calculating the point estimator $\hat{\theta}^{w}_{CE}$ as described above. We performed 1000 repetitions of permutations to obtain the null distribution of the estimator $\hat{\theta}^{w}_{CE}$. The corresponding `R` code is provided as Supporting Information to ease application in practice.

Table 3 shows the average effect estimates over all 1000 random samples and the estimated power values for the weighted all-cause hazard ratio and the proportion in favor of treatment approaches.

It can be seen that $\hat{\theta}_{CE}$, $\hat{\theta}_{EP_1}$, and $\hat{\theta}_{EP_2}$ are reasonably close to the true underlying hazard ratios. The deviation of $\hat{\theta}_{EP_1}$ to θ_{EP_1} can be explained by the underlying small group-specific hazards, which increase the standard deviation of the estimates. The weighted all-cause hazard ratio is estimated in mean as 1.44, which still defines a large negative effect supporting a harmful effect of the intervention. However, the weighted effect is smaller than the unweighted effect, as the contribution of the second component is smaller. Note that the less relevant component EP_2 corresponds to larger underlying hazards than for EP_1. Therefore, EP_2 has a higher impact on the overall effect even after weighting. In contrast, the estimated mean proportion in favor of treatment is given by 0.14 in the scenario with the smaller follow-up duration. This defines a clear effect in favor of the intervention group. Here, the large negative effect for EP_2 has only a minor influence as the majority of pairwise comparisons is either uninformative or favors the intervention group. This is because a comparison with respect to the component of secondary relevance is only performed if no decision with respect to the first component is possible. For the scenario with a longer follow up duration the estimated mean proportion in favor of treatment is given by 0.21, which is an even larger effect in favor of the intervention whereas all other effect measures remain basically unchanged. This can be explained by the fact that if the individual follow-up times are longer, then more events of type EP_1 are observed and thus more pairwise comparisons are based on considering EP_1. This illustrates the high sensitivity of the proportion in favor of treatment on the follow-up distribution. Note that in this simulation setting, exponentially distributed event times were considered meaning that the hazards are constant in time. Therefore, the dependence of the proportion of favor in treatment on the observational time is an unintuitive property, which makes its interpretation quite difficult.

6 | CLINICAL TRIAL EXAMPLES

We will now investigate the performance of the weighted all-cause hazard ratio approach in comparison to the common all-cause hazard ratio which is assessed by the logrank test for the situation of the CAPRICORN Trial introduced earlier.[12] We again consider the following hazard assumptions assuming exponentially distributed event times which are motivated by the original observed and reported results[13]

$$\theta_{CE} = \frac{0.0276}{0.0299} = 0.92,$$
$$\theta_{Death} = \frac{0.008}{0.0104} = 0.769,$$
$$\theta_{Hospital} = \frac{0.0196}{0.0195} = 1.005.$$

To apply the weighted all-cause hazard ratio, the following weights are used as motivated above in Section 4.3

$$w_{Death} = 0.9, w_{Hospital} = 0.1.$$

To assess the robustness of the results under slightly different weights, we additionally investigate $w_{Death} = 0.8$, $w_{Hospital} = 0.2$, and $w_{Death} = 0.7, w_{Hospital} = 0.3$. To directly compare the power of the permutation test to the power of the Cox-model, we also consider $w_{Death} = 0.5, w_{Hospital} = 0.5$. Assuming exponentially distributed event times, the corresponding weighted all-cause hazard ratios can be directly deduced and are provided in Table 4. Note that as we assume exponentially distributed event times, the different effect are reported as constant parameters and the averaged weighted all-cause hazard ratios were omitted.

For equal weights across the components (Table 4, line 1), the weighted all-cause hazard ratio is equivalent to the standard all-cause hazard ratio by construction. From lines 2 to 4, it can be seen that the weighted all-cause hazard ratio is reasonably robust against small changes in the weights. The weighted all-cause hazard ratio shows a larger effect for increasing weight w_{Death}. This is because the component "death" shows a relatively large effect whereas the effect for the component "cardiovascular hospital admission" slightly points in the adverse direction.

To investigate the performance of the corresponding point estimator proposed in (11) and of the permutation test, which is used to assess the underlying test hypotheses provided in (14), we additionally performed a simulation study using the software `R`. For the hazard assumptions specified above, we generated 10 000 random samples with 980 patients per group. Thereby, uniform patient accrual was simulated within the first 24 months with an additional minimal follow-up of 3 months. For each sample, the corresponding hazard estimates within the two groups were obtained by fitting Weibull-models with scale parameters fixed to 1 using the function `survreg` of the `R` package `survival`. The plug-in estimator $\hat{\theta}^w_{CE}$ was obtained by substituting the resulting hazard estimates. The permutation test was performed by randomly reassigning the group affiliations to the original complete data set and subsequently calculating the point estimator $\hat{\theta}^w_{CE}$ as described above. We performed 1000 repetitions of permutations to obtain the null distribution of the estimator $\hat{\theta}^w_{CE}$. The corresponding `R` code is provided as Supporting Information to ease application in practice.

Table 5 shows the average effect estimates over all 10 000 random samples and the estimated power values of the permutation test and the corresponding Cox-models.

Generally, on average the different effect estimators approximate the true parameters provided in Table 4 very well. When assigning equal weights to the components (line 1), the estimator for the weighted all-cause hazard ratio and for the standard all-cause hazard ratio estimate the same parameter. The estimates are indeed nearly the same. The power of the permutation test corresponding to the weighted all-cause hazard ratio is very close to the power of the corresponding

TABLE 4 Weighted all-cause hazard ratio for the CAPRICORN trial

w_{Death}	$w_{Hospital}$	θ^w_{CE}	θ_{CE}	θ_{EP_1}	θ_{EP_2}
0.5	0.5	0.92	0.92	0.769	1.005
0.7	0.3	0.874	0.92	0.769	1.005
0.8	0.2	0.845	0.92	0.769	1.005
0.9	0.1	0.810	0.92	0.769	1.005

TABLE 5 Simulated weighted all-cause hazard ratio and power based on the results of the CAPRICORN Trial

w_{Death}	$w_{Hospital}$	$\hat{\theta}^w_{CE}$ (Power)	$\hat{\theta}_{CE}$ (Power)	$\hat{\theta}_{Death}$ (Power)	$\hat{\theta}_{Hospital}$(Power)
0.5	0.5	0.926 (0.177)	0.926 (0.174)	0.775 (0.482)	1.010 (0.021)
0.7	0.3	0.877 (0.366)	0.926 (0.175)	0.775 (0.482)	1.010 (0.021)
0.8	0.2	0.848 (0.436)	0.926 (0.175)	0.775 (0.482)	1.010 (0.021)
0.9	0.1	0.814 (0.467)	0.926 (0.175)	0.775 (0.482)	1.010 (0.021)

Cox-model. The magnitude of the power, however, is low as the component "cardiovascular hospital admission" shows a slight adverse effect. As a consequence, the combined effect is close to 1.

With increasing weight w_{Death} for the component "death," the power of the permutation test increases and exceeds the power of the Cox-model. This is due to the fact that the component "death" corresponds to a larger effect whereas the component "cardiovascular hospital admission" shows a small adverse effect in terms of the corresponding cause-specific hazard ratios, as pointed out above.

Generally, the weighted approach with reasonably chosen weights $w_{Death} > w_{Hospital} > 0$ might be preferred over the standard unweighted Cox-model for the all-cause hazard ratio as a) it eases interpretation by taking into account the different levels of relevance of the components and b) provides a power advantage for the specific parameter setting of the CAPRICORN Trial. Whereas an interpretation benefit is given irrespective of the specific clinical trial scenario, the power of the weighted approach depends on the underlying cause-specific hazards. There also exist situations, where the weighted approach will have lower power compared to the standard Cox-model, eg, in the common situation when the most relevant endpoint corresponds to the most rare event. To illustrate this situation, an additional clinical trial example is investigated in the following. Note, however, that the primary intention of the weighted approach is to improve the interpretation of the effect measure. Therefore, it is desirable that a weighted effect is shrunken when the most relevant component only corresponds to a small effect and/or is based on a small number of events.

The RENAAL Trial was designed as a randomized, double-blind, placebo-controlled clinical trial with one interim analysis to assess whether the angiotensin-II-receptor anatagonist losartan shows a therapeutic benefit for nephropathy patients with type 2 diabetes.[25,26] The primary endpoint was a composite time-to-first-event endpoint where the event types were given as "death," "end-stage renal disease," or "doubling in the baseline serum creatinine concentration." Patients were randomized in a 1:1 allocation to receive either losartan or placebo. Recruitment duration was planned to require 2 years, and the minimal patients' follow-up duration was chosen to be 3.5 years. The sample size calculation was based on detecting a relative risk reduction of 0.2 in the 5-year cumulative event rate of the composite endpoint from 0.580 in the placebo group to 0.464 in the losartan group with 0.95 power at a global one-sided significance level of 0.025, where the adjusted local significance level of the final analysis was 0.024. Assuming exponentially distributed event times, the above effect corresponds to a hazard ratio of 0.72 (intervention versus control). The total sample size of 1513 patients included a considerable number of patients due to a recruitment overrun.

The results of the RENAAL Trial where not provided by means of the corresponding hazard ratios, which would have been the appropriate effect measure, but by means of the absolute and relative frequencies which ignores the censoring distribution and issues of competing risks. Therefore, the published results should be interpreted with care. However, the published results provide sufficient information to discuss general problems regarding the composite primary endpoint. A number of 975 patients was recruited to the losartan group and 984 patients to the placebo group. The primary composite endpoint was reached by 327 patients in the losartan group (0.435) as compared to 359 patients in the placebo group (0.471), which corresponds to a relevant positive treatment effect. A detailed look into the single component effects reveals a relevantly lower risk in the losartan group for the components "doubling in the baseline serum creatinine concentration" (162 events corresponding to an event rate of 0.216 in the losartan group versus 198 corresponding to a rate of 0.260 in the control group) and "end-stage renal disease" (147 events corresponding to an event rate of 0.196 in the losartan group versus 194 events corresponding to an event rate of 0.255 in the control group), whereas in the particular harmful component "death" a small adverse effect was observed (158 events corresponding to an event rate of 0.210 in the losartan group versus 155 events corresponding to an event rate of 0.203 in the control group). From these results, we again deduced the following underlying hazards and hazard ratios assuming exponentially distributed event times

$$\begin{aligned}
\theta_{CE} &= \frac{0.126}{0.156} = 0.808, \\
\theta_{Death} &= \frac{0.021}{0.020} = 1.05, \\
\theta_{EndStage} &= \frac{0.054}{0.071} = 0.761, \\
\theta_{Doubl} &= \frac{0.051}{0.065} = 0.785.
\end{aligned}$$

It can be seen that the most relevant component "death" shows a small adverse effect. Thus, when assigning a higher weight to the most relevant component death, the combined effect will be shrunken. This can be a situation where a "benefit-risk assessment" to determine the weights as described in Section 4.3 might be appropriate. The following weights

TABLE 6 Simulated weighted all-cause hazard ratio and power based on the results of the RENAAL Trial

w_{Death}	$w_{EndStage}$	w_{Doubl}	$\hat{\theta}_{CE}$ (Power)	$\hat{\theta}^{w}_{CE}$ (Power)	$\hat{\theta}_{Death}$ (Power)	$\hat{\theta}_{EndStage}$(Power)	$\hat{\theta}_{Doubl}$ (Power)
0.5	0.4	0.1	0.811 (0.806)	0.833 (0.585)	1.076 (0.014)	0.766 (0.667)	0.790 (0.531)
0.45	0.35	0.2	0.810 (0.804)	0.825 (0.698)	1.071 (0.012)	0.766 (0.665)	0.789 (0.540)

are chosen as starting values to investigate the weighted all-cause hazard ratio.

$$w_{Death} = 0.5, \quad w_{EndStage} = 0.4, \quad w_{Doubl} = 0.1.$$

As before, Table 6 shows the average effect estimates over 10 000 random samples as well as the estimated power values of the permutation test and the corresponding Cox-models.

It can be seen that the power loss using the weighted all-cause hazard ratio is considerable. While the original all-cause hazard ratio refers to a power value of 0.806 the power for the weighted all-cause hazard ratio is only 0.585. This might impose the question how large the weight for the component "death" can be chosen without too much loss in power. Line 2 of Table 6 shows a constellation of weights which is similar to the starting values but yielding a power of approximately 0.7, which might still seem acceptable. These weights may thus be seen as an appropriate choice under "risk-benefit" considerations. Note, again, that the rationale to introduce weights is not to increase the power but to improve the clinical interpretation. While the CAPRICORN Trail illustrates a situation where the weights result in a power increase, the RENAAL Trial shows that the gain in interpretation can also come along with a loss of power.

7 | DISCUSSION

In this work, we presented a weighted all-cause hazard ratio as an alternative effect measure to the standard all-cause hazard ratio to assess a composite time-to-first-event endpoint. The weights must be prespecified in the planning stage to reflect the different levels of clinical relevance of the components. In contrast to other approaches proposed in the past,[20] the weights are directly assigned to the individual cause-specific hazards and not to the (log-transformed) cause-specific hazard ratios. By this, our approach defines a natural extension of the standard all-cause hazard ratio. In particular, both effect measures are equivalent if equal weights are chosen across components. The new effect measure is a time-dependent measure, which can either be reported for a specific follow-up time point or averaged over the observational period. The theoretical investigations and the simulation study provided in Sections 5 and 6, respectively, illustrate the advantage of our new approach: The weighted all-cause hazard ratio is more strongly influenced by components with high clinical relevance whereas components of low relevance have a low impact. As a consequence, the weighted all-cause hazard ratio defines a clinically more relevant effect measure. This, however, does not imply anything about the magnitude of the effect. The weighted all-cause hazard ratio shows a higher effect than the standard all-cause hazard ratio whenever the components of high clinical relevance show large effects. In contrast, the weighted all-cause hazard ratio will show a smaller effect than the standard hazard ratio if treatment effects are only observed for components with low clinical relevance. The intention to use the weighted all-cause hazard ratio is thus to improve the interpretation but not necessarily to increase the power of the trial. Generally, an effect measure should primarily be chosen from a clinical perspective to provide a meaningful quantification of the treatment effect.

The magnitude of the underlying cause-specific hazards additionally determines the size of the weighted all-cause hazard ratio. If a single component shows a large effect (in terms of the cause-specific hazard ratio) but the underlying cause-specific hazards are small, then this effect relies on only a few events. Therefore, it is desirable that small cause-specific hazards have a lower impact on the weighted effect than large cause-specific hazards. As pointed out in Section 5, the impact of the magnitude of the underlying hazards on the weighted all-cause hazard ratio is an important difference to the approach proposed by Lachin and Bebu[20] where the magnitude of the individual hazards has no influence but only the magnitude of the individual hazard ratios.

Our new approach can easily be extended to take account of covariates by substituting the covariate-adjusted estimates of the cause-specific hazards from the corresponding survival model. Note, however, that in the presence of covariates, the standard all-cause hazard ratio estimated from the Cox-model and the weighted all-cause hazard ratio might differ even in the case of equal weights across components, as covariates are modeled differently in both approaches. A more detailed investigation of this issue is topic of an ongoing work. Moreover, multiple events per patient can be considered by

additionally including weighted hazards for the transition between a first and a second event or between any subsequent events.

A potential drawback of our method is that a parametric survival model is required to identify the underlying cause-specific hazards. However, considering the large number of possible parametric survival models, this is not a strong restriction with respect to flexibility. However, the choice of an adequate parametric survival model in the planning stage defines a major challenge. Assessing the robustness of the estimators under misspecifications of the underlying survival model is one of the aims of our ongoing research.To partly address this issue, we additionally presented a nonparametric estimator for the weighted all-cause hazard ratio for the special case of equal baseline-hazards across the components. It may be seen as another disadvantage of our approach that the distribution of the test statistic cannot directly be deduced and hence nonparametric conditional tests such as the permutation test have to be applied for hypotheses testing. We have implemented the permutation test for the weighted all-cause hazard ratio in the software `R` and provide the source code as Supporting Information. By this, application of our method in practice can be done with minimal effort for each parametric model where hazard functions are estimable. Moreover, the permutation test offers the advantage that it remains valid under misspecifications of the underlying survival model. Even if the point estimator is biased, the permutation test still defines a valid test strategy for this slightly modified effect measure.

Generally, weighted approaches might be also defined using additive survival models. Using a weighted average of component effect measures derived from an additive model has the advantage that the underlying distribution can be easily deduced and thus a parametric test can be defined. Moreover, using an additive model would not require the specification of parametric survival models to identify the individual hazards. An important drawback of this appealing alternative is, however, that additive models are rarely met in practice and a direct comparison to the standard Cox-model for the all-cause hazard ratio is not possible. Still, this approach is an interesting topic for future research.

We additionally provided a small simulation study to illustrate the difference between the weighted all-cause hazard ratio and the commonly cited proportion in favor of treatment proposed by Buyse.[16] The results show that the proportion in favor of treatment strongly depends on the individual follow-up distribution and therefore does not properly defines a weighting approach with predefined relevance weights. These results are in common with the conclusions from other authors,[17,18] which support the recommendation that the proportion in favor of treatment might not be a good effect measure when applied to several time-to-event endpoints.

Although composite endpoints most often correspond to time-to-event variables, there also exist situations where a binary composite endpoint is considered. The presented approach might be generalized to define a corresponding weighted effect measure for binary composite endpoints. A related weighting approach for binary composite endpoints was also proposed by Duc and Wolbers.[14]

ACKNOWLEDGEMENTS

Geraldine Rauch was supported by the German Research Foundation (grant number: RA 2347/1-2). We thank Jan Beyersmann for the helpful comments that considerably helped to improve our manuscript.

ORCID

Geraldine Rauch http://orcid.org/0000-0002-2451-1660
Kevin Kunzmann http://orcid.org/0000-0002-1140-7143
Meinhard Kieser http://orcid.org/0000-0003-2402-4333

REFERENCES

1. Lubsen J, Kirwan BA. Combined endpoints: can we use them? *Stat Med*. 2002;21:2959-29170.
2. ICH guideline. Statistical principles for clinical trials (E9). 1999. http://www.fda.gov/downloads/drugs/guidancecomplianceregulatory/information/guidances/ucm073137.pdf. Accessed November 22, 2017.
3. Rauch G, Beyersmann J. Planning and evaluating clinical trials with composite time-to-first-event endpoints in a competing risk framework. *Stat Med*. 2013;32:3595-3608.
4. Bethel MA, Holman R, Haffner SM, et al. Determining the most appropriate components for a composite clinical trial outcome. *Am Heart J*. 2008;156:633-640.
5. Freemantle N, Calvert M. Composite and surrogate outcomes in randomised controlled trials. *Br Med J*. 2007;334:756-757.

6. Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. Composite outcomes in randomized trials—greater precision but with greater uncertainty?*J Am Med Assoc*. 2003;289:756-757.
7. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. General Methods - Version 4.2. 2015. https://www.iqwig.de/download/IQWiG_Methoden_Version_4-2.pdf. Accessed November 22, 2017.
8. CPMP Guideline. Points to consider on multiplicity issues in clinical trials. 2002. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003640.pdf. Accessed November 22, 2017.
9. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multistate models. *Stat Med*. 2007;26:2389-2430.
10. Andersen PK, Keiding N. Multi-state models for event history analysis. *Stat Methods Med Res*. 2002;11:91-115.
11. Eulenburg C, Mahner S, Woelber L, Wegscheider K. A systematic model specification procedure for an illness-death model without recovery. *PloS one*. 2015;e0123489:10.
12. Dargie H, the CAPRICORN Steering Committee. Design and methodology of the CAPRICORN trial—a randomised double blind placebo controlled study of the impact of carvedilol on morbidity and mortality in patients with left ventricular dysfunction after myocardial infarction. *Eur J Heart Fail*. 2014;21:74-80.
13. Dargie H, the CAPRICORN Investigators. Effect of carvedilol on outcome after myocardial infarction in patients with left-ventricular dysfunction: The CAPRICORN randomised trial. *Lancet*. 2001;357:1385-1390.
14. Duc AN, Wolbers M. Weighted analysis of composite endpoints with simultaneous inference for flexible weight constraints. *Stat Med*. 2017;36:442-454.
15. Pocock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *Eur Heart J*. 2012;33:176-182.
16. Buyse M. Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Stat Med*. 2010;29:3245-3257.
17. Bebu I, Lachin JM. Large sample inference of a win ratio analysis of a composite outcome base don prioritized outcomes. *Biostatistics*. 2016;17:178-187.
18. Rauch G, Jahn–Eimermacher A, Brannath W, Kieser M. Opportunities and challenges of combined effect measures based on prioritized outcomes. *Stat Med*. 2014;33:1104-1120.
19. Péron J, Buyse M, Ozenne B, Roche L, Roy P. An extension of generalized pairwise comparisons for prioritized outcomes in the presence of censoring. *Stat Methods Med Res*. 2016. https://doi.org/10.1177/0962280216658320
20. Lachin JM, Bebu I. Application of the WeiLachin multivariate one-directional test to multiple event-time outcomes. *Clin Trials*. 2015;12:627-633. https://doi.org/10.1177/1740774515601027
21. Andersen P, Borgan Ø, Gill R, Keiding N. *Statistical Models Based on Counting Processes*. New York: Springer; 1993.
22. Chow S, Shao J, Wang H. *Sample Size Calculations in Clinical Research*. Boca Raton: Chapman & Hall; 2008.
23. Schoenfeld D. The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrica*. 1981;68:316-319.
24. Lehmann EL. *Elements of Large-Sample Theory*. New York: Springer; 1999.
25. Brenner MB, Cooper ME, de Zeeuw Dc, et al. The losartan renal protection study—rationale, study design and baseline characteristics of RENAAL (Reduction of endpoints in NIDDM with the angiotensin II antagonist losartan). *J Renin Angiotensin Aldosterone Syst*. 2000;1:328-335.
26. Barry M, Brenner BM, Cooper ME, et al. Effects of losartan on renal and cardiovascular outcomes in patients with type 2 diabetes and nephropathy. *N Engl J Med*. 2001;345:861-869.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Rauch G, Kunzmann K, Kieser M, Wegscheider K, König J, Eulenburg C. A weighted combined effect measure for the analysis of a composite time-to-first-event endpoint with components of different clinical relevance. *Statistics in Medicine*. 2018;37:749–767. https://doi.org/10.1002/sim.7531